ISSN: 2348 9510
**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 1, Issue 5, August 2014**

# Clustering In Data Mining : A Brief Review

**Meenu Sharma**
Mtech scholar,
Department of computer science,SD Bansal college of tehnology ,Indore.

## Abstract

Data analysis plays an important role in understanding various phenomena.Clustering has got a significance attention in data analysis,image recognition,control process,data management,data mining etc. Due a enormous increment in the assets of computer and communication technology.Cluster analysis aims at identifying groups of similar objects and, therefore helps to discover distribution of patterns and interesting correlations in large datasets.This review paper acts as a catalyst in the initial study of the various researchers who directly or indirectly deals with clustering in their research work.In this paper,a comprehensive study of clustering is done along with its all techniques and a simple comparison of them,so that it is easy for someone to pick a specific method as per suitable to the working environment.

## Introduction

Cluster analysis is a technique which discovers the substructure of a data set by dividing it into several clusters.The term “clustering” is used in several research communities to describe methods for grouping of unlabeled data[1]. It is a useful and basic tool for data analysis in which similar item are bunched into one partition and likewise different items are clubbed in different parts/partition[2].Clustering is also a firmamental operation in data mining.Data mining is the process of analyzing a specific data from various perspectives and then concluding it into a value added information.The steps involves in data mining is depicted in fig.1.

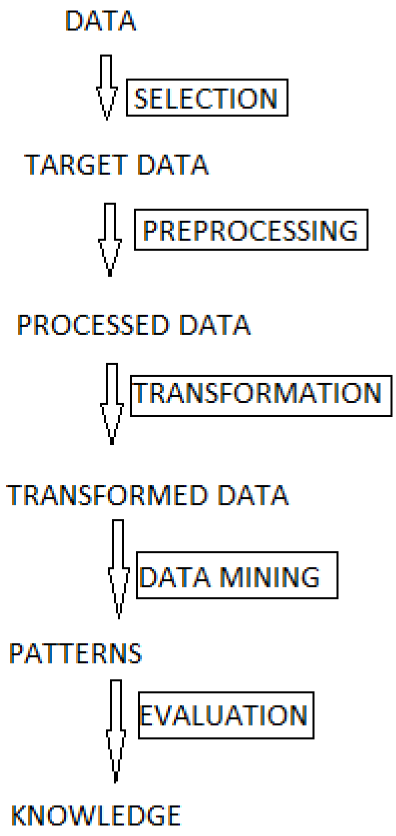


Fig. 1 Steps of Data Mining Process.

Data mining involves the anomaly detection, association rule learning, classification, regression, summarization and clustering.In this paper, clustering,a integral step of data mining is analysis as per the past research work done on it.In data mining the data is mined using two learning approaches i.e. supervised learning or unsupervised clustering[3].In supervised learning,the correct result are known and are given to the model during the training process.This methods is accurate and fast as compared to unsupervised learning techniques.Some examples of this methods are neural network,multilayer perception,decision trees etc.However,in the unsupervised learning,the model is not provided with the correct results during the learning.K-means,self organizing maps,distances and normalization comes under the unsupervised learning.Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.In other words it is a undirected data mining process.Different types of clusters are given in the table 1.

| NAME OF CLUSTER | BRIEF DESCRIPTION |
|---|---|
| Well Separated Clusters | A cluster in which all the points in a particular cluster is same or nearest to other points of the cluster as compared to the other points which are not in the cluster is called well separated cluster. |
| Centre Based clusters | A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the “center” of a cluster, than to the center of any other cluster. The center of a cluster is often a centroid. |
| **Contiguous clusters** | A cluster is a set of points so that a point in a cluster is nearest (or more similar) to one or more other points in the cluster as compared to any point that is not in the cluster. |
| **Density-based clusters** | Cluster which are having same density of points in clusters differentiated as low density and high density, are called density based clusters. |
| **Shared Property or Conceptual Clusters** | Clusters that shares common property or concepts are called shared or conceptual clusters. |

Table. 1[3] Different types of Clusters.

## Different Stages Of Clustering

Different stages in clustering along with the feedback loop are given in the below figure:

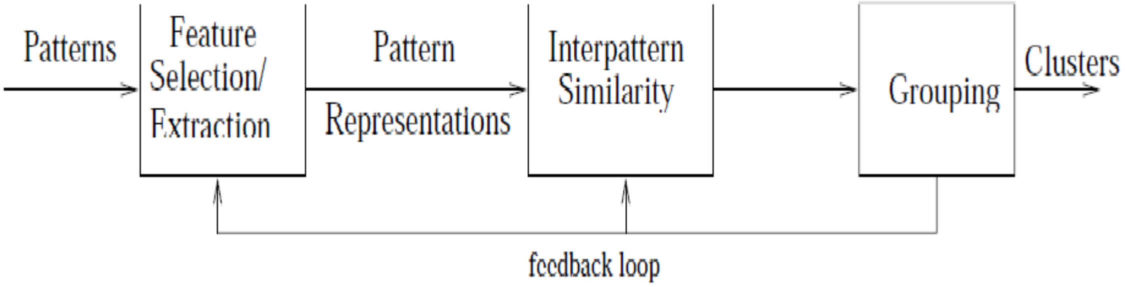


Fig. 2 [1] Different Stages Involved In Clustering.

In pattern representation,different types and number of the classes,patterns and scale of the features available in the clustering algorithms are represented.In Feature selection,identification of the effective subset of the original features to use in the clustering is done while in feature extraction,transformations of the input features is done so as to produce new salient features.Inter-pattern similarity is measured by a distance function defined on pairs of patterns.Anderberg[4], Jain and Dubes[5] and Diday and Simon[6] uses different types of distance measures.Euclidean is one such distance measure used to reflect dissimilarity between two patterns.The grouping step can be performed wither in hard way or in fuzzy.The former implies a partition of the data into groups while later implies that each pattern has a variable degree of membership in each of the output clusters[1].The selection of grouping methods depends upon the type of clustering tobe used further.In Data abstraction the data set is represent in a well mannered and readable form.Feedback loop is used to obtain continuos improvement in the data set.

## Different Techniques Of The Clustering

There are many classification suggested in literature.The broad classification as given by Jain and Dubes [7] is Hierarchical and partitional clustering.All other methods of clustering is the under this two broad types.Recent literature reveals new techniques of clustering such that density based algorithms, grid based algorithms, ANFIS(Adaptive Network Based fuzzy interference system) and FCM(Fuzzy c-means clustering method). A brief description of all methods are given below.

## Hierarchical Clustering

This method creates a hierarchical decomposition of the given set of data objects. The tree of clusters so formed named as dendrograms. Every cluster node contains child clusters, sibling clusters partition the points covered by their common parent[8]. In hierarchical clustering,the number of items are equal to the number of clusters(say n). The pairs which are closest to each other are merged into single cluster.After this measurement of the distance between new cluster and each of old clusters.Repeating of the steps is done until all items are clustered into m no. of clusters.This can be done in two ways:

A) Bottom Up hierarchical clustering: In this type of clustering,a parent cluster in divided into fragmented cluster which in turn again splits into clusters.The methods starts with merging cluster into larger one until all the objects are in a single cluster or certain termination condition is satisfied.This method is also called as agglomerative clustering.During agglomeration,the closest cluster are merged into a single cluster[9].

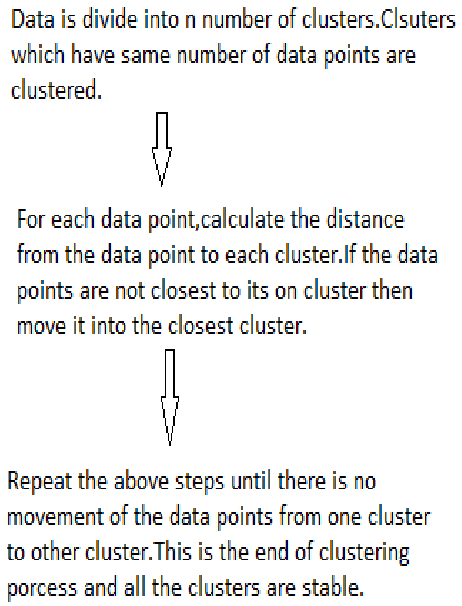
B) Top Down hierarchical clustering: Unlike agglomerative clustering,this method starts with a single cluster containing all objects and then progressively splits which ultimately clustered until only individual clusters remain[10].This method is also called as Divisive clustering which is not commonly used.

Hierarchical clustering enjoys the benefits of flexibility,ease of handling and applicability to any attribute type but it suffers from a serious demerits that once a step (merge or split) is done, it can never be undone.

## Partitioning Clustering

Methods like k-mean, Bisecting K Means Method, Medoids Method, PAM (Partitioning Around Medoids), CLARA (Clustering LARGE Applications) and the Probabilistic Clustering are comes under partitioning clustering.The name itself suggests that the data is divided into number of subsets.Since it is not computationally possible to check the all possible subset of the systems available that is the reason this methods can be used to clustered large data.to overcome such limitation of checking,this methods uses statistical method to assign rank values to the cluster categorical data.This data One such method is k-means method whose steps are depicted if fig.3.This method which find mutual exclusive clusters of spherical method.The categorial data so obtain from the statical method has been convert into numeric by assigning rank value to them[11].

This method is efficient in processing large data and always terminates with a optimum results with clusters of convex shape.



Data is divide into n number of clusters.Clsuters which have same number of data points are clustered.

For each data point,calculate the distance from the data point to each cluster.If the data points are not closest to its on cluster then move it into the closest cluster.

Repeat the above steps until there is no movement of the data points from one cluster to other cluster.This is the end of clustering porcess and all the clusters are stable.

Fig. 3 Steps involved in implementation of k-means technique.

## Fuzzy C-Means Clustering Methods

All the above mentioned methods have cluster boundaries which are defined for data and data elements are sharp but in real problems the features or attributes are not that much sharp because they have some potential to be a part of some other class to a particular extent[2].The use of Fuzzy logic theory overcome this limitation.Fuzzy logic takes into account the degree of uncertainty of samples belonging to each classes and also their relationships,thus they reflect the real world situation.Also the patterns formed by partitions as discussed by the previous mentioned methods have relation with one and only one cluster and thus the cluster so formed have are disjoint.Zadeh[12] has concluded that fuzzy clustering has a feature which is supported by a membership function.One of the best fuzzy methods is fuzzy c-means (FCM) algorithm which is depicted by Bezdek[13,14] in his papers in 1974 and 1981 respectively.FCM method works on the optimization of a specific cost function, and it operates well when the clusters are compact or isotropic.Some advantages of FCM is great ability to detect hyper volume clusters and thin shells - curves and surfaces.Its also shows relationship between patterns of different clusters.A detailed algorithm is very well explained in the review paper by Jain et. al[5].

One can read better on fuzzy clustering in the famous book by Bezdek[14].

## Anfis : Adaptive Neuro Fuzzy Inference System

On 1993,ROGER Jang developed the ANFIS techniques that over comes the short comings of ANNs and  fuzzy systems. ANFIS uses neural and fuzzy logic approaches at same time to combine the advantages of each method to achieves better performance. The main motto of the ANFIS technique is to merge the remarkable features of the Fuzzy systems and neural networks.It reduce the optimization search space by highlighting the prior knowledge into a set of constraints,which is one of the advantage of fuzzy system.Moreover,it also automate FC parametric tuning by adapting the back propagation to structured network.ANFIS finds its best application in controllers,which is used to automate FC tuning and modelers,which are sued to explain past data and predict future behaviour. Fuzzy part of ANFIS is constructed by mean of input and output variables , membership functions ,fuzzy rules and inference method.

Membership functions are the functions that defines the fuzzy sets .the fuzzy rules have a form of if-then rules and define how the output must be for a specific value of membership of its inputs.The basic steps involved in the implementation of the ANFIS are shown in the fig.5

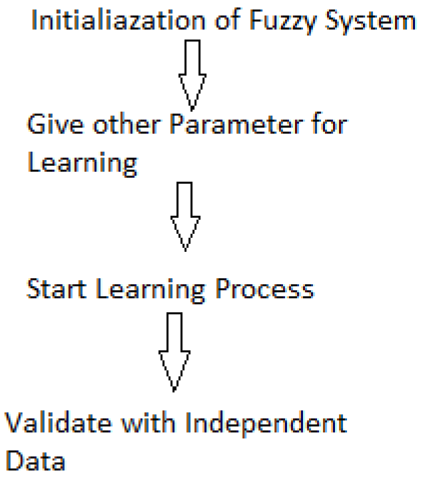

Initializaztion of Fuzzy System
Give other Parameter for Learning
Start Learning Process
Validate with Independent Data

Fig.5 The various steps of implementing the AFNIS techniques.

Initialization of fuzzy system takes place with the help of the GENIS 1 AND GENIS 2 commands.After then number of iterations and error in the form of tolerance is given as input.ANFIS commands is given then and waits for the results which meets our tolerance level.The result so obtained is then validate with independent data.

The Fuzzy controllers system or Fuzzy Inference system as suggested by [15] is given in figure 4.

Fuzzy inference system (FIS) is also known as fuzzy rule base system , fuzzy model, fuzzy associative memories or fuzzy controller.

Basically FIS is composed of 5 functional blocks:

. A fuzzyfication interface which transforms crisp inputs into degrees of match with linguistic rule values.

. A rule base containing number of fuzzy if then rules.

. A data base which defines the membership functions of the fuzzy set used in the fuzzy rules.

. A defuzzyfication interface which transforms the fuzzy results into a crisp outputs.

DATABASE and RULEBASE are jointly referred to as a knowledge base.

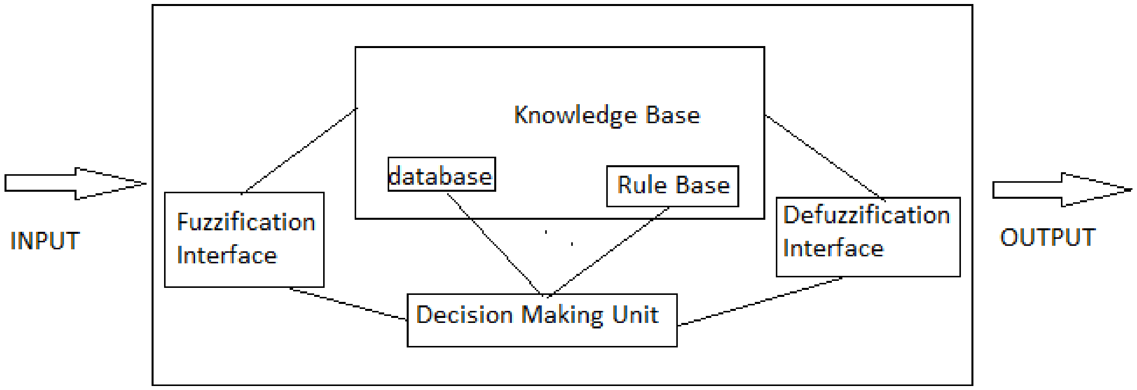
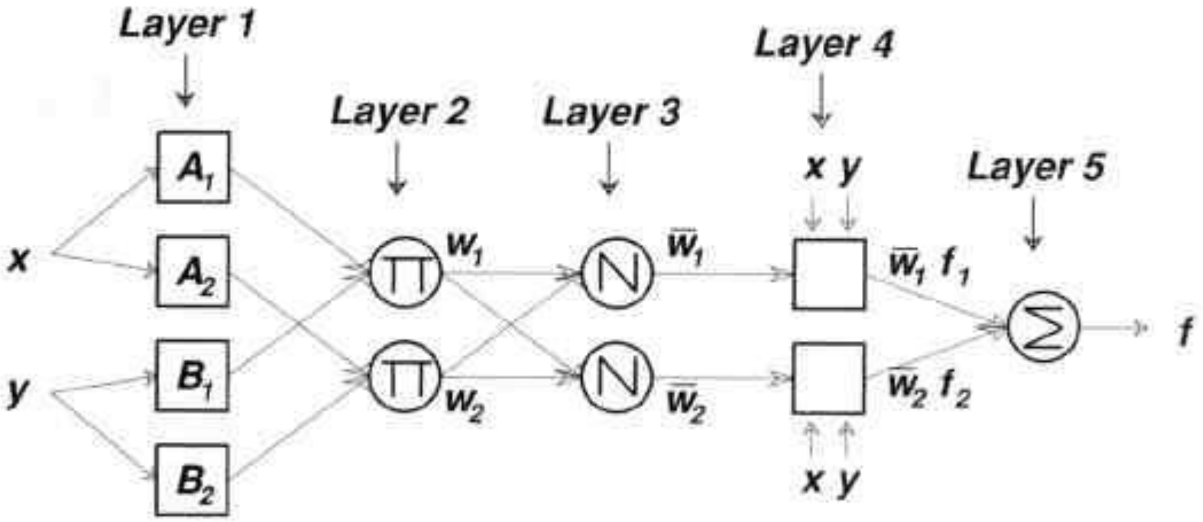


Fig.4 [15] The Fuzzy Interface System.

5 layer architecture of ANFIS:



The first layer of the structure is called fuzzification.

In the second layer, the weight of each rule has to be computed by means of a fuzzy AND operation.

In the layer 3, it is made the normalization of the values.

in the layer 4 the defuzzification process.

Finally in layer 5, the overall output of the system is obtained .

## Conclusion

In this paper,we have successfully represent all the techniques along with their implementation methods and relative merits and demerits among each other.We hope this paper will help all the scholars whose choose clustering as their research.However,one can also refer detailed literature available on the various methods mentioned in this paper.ANFIS and FCM methods respond better to real life situation as compared to other methods.Depending upon the situation and tolerance level desire,the selection of the methods is done,

## References

[1] A.K. JAIN,M.N. MURTY&P.J. FLYNN,”Data Clustering: A Review ”,ACM Computer Survey, vol. 31, no. 3, pp. 264–323, 1999.

[2] Farhat Roohi,” NEURO FUZZY APPROACH TO DATA CLUSTERING: A FRAMEWORK FOR ANALYSIS”, European Scientific Journal March 2013 edition vol.9, No.9 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431.

[3]Amandeep Kaur Mann ,Navneet Kaur ,”Survey Paper on Clustering Techniques ”,International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[4] Anderberg, “ Cluster Analysis for Applications.”Academic Press, Inc., New York, NY.

[5] DUBES, R. C.& JAIN, A. K. ,”Clustering techniques: The user’s dilemma”, Pattern Recogn. *8*, 247–260.

[6] DIDAY, E. AND SIMON, J. C.,” Clustering analysis. In Digital Pattern Recognition”K. S. Fu, Ed. Springer-Verlag, Secaucus, NJ, 47–94.

[7] JAIN, A. K. AND DUBES, R. C. ,” Algorithms for Clustering Data”,Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ.

[8] Aastha Joshi, Rajneet Kaur ,” A Review: Comparative Study of Various Clustering Techniques in Data Mining “ Volume 3, Issue 3, March 2013,International Journal of Advanced Research in Computer Science and Software Engineering.

[9] Improved Outcome Software, Agglomerative Hierarchical Clustering Overview. Retrieved from: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Agglomerative_Hierarchical_Clustering_Overview.htm [Accessed 22/02/2013].

[10] Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd ed, 443-491.

[11] Patnaik, Sovan Kumar, Soumya Sahoo, and Dillip Kumar Swain, “Clustering of Categorical Data by Assigning Rank through Statistical Approach,” International Journal of Computer Applications 43.2: 1-3, 2012.

[12] ZADEH, L. A.,” Fuzzy sets” Inf. Control *1965*, *8*,338–353.

[13] J. Bezdek, Fuzzy mathematics in pattern classification, Ph.D. thesis, Ithaca, NY: Cornell University, 1974.

[14] BEZDEK, J. C.,” Pattern Recognition With fuzzy Objective Function Algorithms” Plenum Press, New York, NY.1981.

[15] Shing, Roger and Jhang,“AFNIS- Adaptive Network Based Fuzzy Inference System”,IEEE transaction On systems,man and cybernetics, vol 23, no. 3, JUNE 1993.